質問応答の強化学習による抽象型要約の精度向上

高塚 雅人　　小林 哲則　　林 良彦
早稲田大学 理工学術院
takatsuka@pcl.cs.waseda.jp

1 はじめに

近年, 抽象型要約生成において強化学習を利用する研究が多く行われている. 一般に抽象型要約に用いられる単語レベルでの最尤推定の loss では, 生成された文全体に対しての loss は計算されないため, 要約文全体を考慮した生成を行うことができない. 強化学習によって, 生成された要約文全体を評価する ROUGE[1] などの微分不可能な指標を最適化することが可能となる. 強化学習の報酬に ROUGE を用いたもの [2] 以外にも要約文に対する質問応答におけるスコアを報酬として用いるもの [3] などがある.

本研究では, 原文章中の重要な情報に対する質問に答えられるような要約には, 原文章中の重要な情報を保存されていることが期待できると考え, Scialom ら [3] と同じく人が生成した要約から質問を生成し, その質問応答のスコアを報酬として強化学習を行う. またその際に, Scialom らの固有名詞を対象とした質問生成を変更して, 名詞を対象とした質問を生成し, その質問に対する正答率を報酬として強化学習を行うことを提案する. これによって固有名詞が要約に出現しないデータでも強化学習を行うことが可能となる. また, 固有名詞と名詞両方の質問を使用することで, 一つの要約に対する質問数が増え, それによってモデルが生成した要約に原文書の重要な情報が保存されているかをより正確に評価できることを期待する.

2 関連研究

一般に要約の評価には ROUGE が用いられるが, ROUGE はモデルが生成した要約と人が作成した要約間の n-gram のマッチングを計算しているだけであり, 要約の読みやすさや流暢性などを測ることはできない. そのためいくつかの研究 [3][2] では, ROUGE のみを強化学習の報酬として用いた場合, 生成された要約の人間による読みやすさの評価が低くなることが報告されている.

ROUGE に代わる指標として, 質問応答の正答率を用いたニュース記事の要約の評価指標に APES[4] がある. APES は機械読解の技術の発展を背景に, 要約が原文書の重要な登場事物に関する質問に回答できる能力を測る指標である. APES では, 人が作成した要約から固有名詞の穴埋め問題を生成し, その問題の正答率でモデルが生成した要約の評価を行う. ニュース記事における要約の人の評価との相関係数が, ROUGE に比べ APES の方が高くなったと報告している.

Scialom ら [3] は, 質問応答によるスコアと ROUGE を足し合わせた値を報酬として用いることを提案している. この研究では, 要約モデルが生成した要約文を参照しながら, QA システムが質問に答え, その質問応答のスコアが向上するように強化学習を行う. Scialom らは, 質問応答のスコアを報酬に加えることで, 人間による要約の読みやすさの評価の下げ幅を抑えつつ, 要約の関連性の評価を向上させることができたと報告している. 質問応答に使用する質問 (穴埋め問題) は, 人が作成した要約内の固有名詞をマスクして生成している.

3 提案手法

3.1 提案モデル

本研究では, 各要約に対する名詞または固有名詞の穴埋め問題を生成し, その正答率を要約モデルの損失関数に追加して強化学習を行う. 提案する要約手法の概要図を図 1 に示す.

各要約に対する穴埋め問題は CNN-Daily-Mail-Reading-Comprehension-Task[5] の手法を参考に作成する. まず, Stanford CoreNLP[1] を使用して, 人が作成した要約の品詞解析を行う. 次に, 品詞解析の結果を用いて, 要約内の名詞と固有名詞を'entity+番号' とマ

[1] https://stanfordnlp.github.io/CoreNLP/

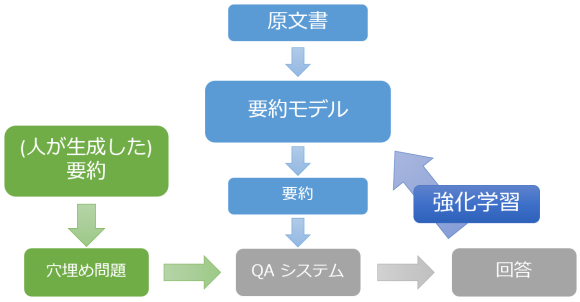

図 1: 提案する要約手法の概要図

スクする. 最後にマスクした単語の中から名詞または固有名詞を一つに選び穴埋め問題を作成する.

3.2 目的関数

学習に使用する loss を定義する. 入力する原文書を $X = [x_0, ..., x_n]$, 人が作成した要約を $Y = [y_0, ..., y_T]$ と置くと, 最尤推定の $loss_{mle}$ は, 以下の式になる.

$$loss_{mle} = -\sum_{t=0}^{T} \log(p(y_t|y_0, \ldots, y_{t-1}, X)) \quad (1)$$

また一般に sequence-to-sequence モデルでは, 出力される単語の繰り返しが問題となるため, 繰り返しを防ぐための coverage loss[6] を導入する. 時刻 t における encoder の LSTM の各出力と decoder の LSTM の最終出力間の attention を a_t とし, coverage vector c_t を

$$c_t = \sum_{t'=0}^{t-1} a_{t'} \quad (2)$$

と置くと, coverage loss は以下の式で定義される.

$$loss_c = \sum_t \sum_i min(a_t^i, c_t^i) \quad (3)$$

最後に強化学習の loss を定義する. 上記の手法で作成した穴埋め問題に QA システムを用いて解答した際の正答率を報酬 r とおく. Self-Critical Training Approach[7] より, モデルの出力からサンプリングした単語系列を $\omega^s = [\omega_0^s, ..., \omega_{T_s}^s]$, テスト時のアルゴリズムで選択した単語系列を $\omega^b = [\omega_0^b, ..., \omega_{T_b}^b]$ とおくと, 強化学習における $loss_{rl}$ は以下の式になる.

$$loss_{rl} = -(r(\omega^s) - r((\omega^b)) \sum_{t=1}^{T_s} \log(p(\omega_t^s|\omega_1^s, \ldots, \omega_{t-1}^s)) \quad (4)$$

最終的な $loss$ は,

$$loss = \lambda_0 loss_{mle} + \lambda_1 loss_c + \lambda_2 loss_{rl} \quad (5)$$

と定義する. $\lambda_0, \lambda_1, \lambda_2$ はハイパーパラメータである.

4 実験

4.1 データセット

CNN-DailyMail dataset(CNN-DM)

ニュース記事 (平均 781 単語) と人が作成した要約 (平均 56 単語) がセットになったデータセット [5]. 学習用のデータが 287,113 ペア, validation 用のデータが 13,368 ペア, テスト用のデータが 11,490 ペア含まれている. See ら [6] が作成した non-anonymized version のデータセットを使用した.

New York Times Annotated Corpus(NYT)

Durrett ら [9] と同様に, NYT 内の 110,540 個のニュース記事 (平均 800 単語) と要約 (平均 46 単語) のペアを抜き出し, 100,834 ペアを学習用に, 9,706 ペアをテスト用に分割した. またテストデータの内から要約長が 50 単語未満のものを取り除き, 3,452 ペアをテストデータとして使用した (NYT50). 学習用のデータの内 4000 個を validation 用のデータとして使用した. 約 2k 個の要約に固有名詞が出現しないデータセットとなっている.

4.2 質問生成

各データセット内の人が作成した要約から, 固有名詞をマスクした穴埋め問題と名詞をマスクした穴埋め問題を作成し, 強化学習に使用した. いくつかのデータには名詞または固有名詞が含まれていなかったため, 学習時にはそれらのデータを取り除いた. CNN-DM データセットで使用する固有名詞の質問は CNN-Daily-Mail-Reading-Comprehension-Task[5] で使われる質問をそのまま使用した. それ以外の質問は 3 節で示した手法を使い, 生成を行った. 生成した質問数, 学習に使用したデータ数を表 1 に示す.

4.3 実験設定

強化学習に用いる QA システムは Danqi ら [10] のモデルを CNN-DM データセットのニュース記事と固

表 2: CNN-DM データセットでの結果

モデル	Acc_{nnp}	Acc_{nn}	Acc_{nnp_nn}	R-1	R-2	R-L
pointer-generator[6]	-	-	-	39.53	17.28	36.38
baseline	45.20	20.71	27.07	39.55	17.43	36.35
+ QA_{nnp}	55.23	-	-	40.26	18.20	36.80
+ QA_{nn}	-	25.96	-	40.56	18.18	37.06
+ QA_{nnp} + QA_{nn}	-	-	30.95	40.68	18.26	37.22

表 3: NYT50 データセットでの結果

モデル	Acc_{nnp}	Acc_{nn}	Acc_{nnp_nn}	R-1	R-2	R-L
pointer-generator[8]	-	-	-	43.71	26.40	-
baseline	27.50	22.67	22.39	42.37	26.40	39.02
+ QA_{nnp}	29.24	-	-	44.46	27.60	41.02
+ QA_{nn}	-	24.43	-	44.81	27.82	41.34
+ QA_{nnp} + QA_{nn}	-	-	23.36	44.66	27.74	41.20

表 1: 生成した質問

データセット	Question Type	データ数	質問数
CNN-DM	固有名詞	287,113	1,259,604
	名詞	286,980	3,392,501
NYT50	固有名詞	98,619	487,237
	名詞	100,517	1,203,241

有名詞の質問で学習させたモデルを用いた. 要約モデルの baseline には, pointer-generator network[6] を使用した. CNN-DM データセットでは入力するニュース記事は 400 単語, NYT50 データセットでは 800 単語に制限をした. また両方のデータセットでモデルが出力する要約は 100 単語に制限をした. 強化学習に使用した質問の設定は以下の三つになる.

- QA_{nnp}
  固有名詞の質問を強化学習に使用する.
- QA_{nn}
  名詞の質問を強化学習に使用する.
- QA_{nnp} + QA_{nn}
  固有名詞と名詞の質問を強化学習に使用する.

式 (5) において $\lambda_0 = 1$, $\lambda_1 = 1$, $\lambda_2 = 0$ とし, baseline モデルを学習させた後, $\lambda_0 = 0.001$, $\lambda_1 = 0.001$, $\lambda_2 = 1$ として強化学習を行った. CNN-DM データセットの評価指標には ROUGE の F 値を使用し, NYT50 データセットの評価指標には, モデルが出力する要約長を人が作成した要約長以下に制限した時の ROUGE の Recall を使用する.

4.4 結果・考察

CNN-DM における実験結果を表 2 に示す. 表の Acc_{nnp} がテストデータにおける固有名詞を対象とした質問の正答率, Acc_{nn} が名詞を対象とした質問の正答率, Acc_{nnp_nn} が固有名詞と名詞両方を対象とした質問の正答率を示す. 表 2 から, baseline の結果と比べ, 強化学習を行った各モデルは, 強化学習に使用した質問の正答率が向上していることがわかる. このことから強化学習は正常に学習できていると推測できる. また固有名詞の質問を強化学習に用いるよりも, 名詞の質問を用いた方が ROUGE が高くなっていることが確認できる. 最も ROUGE が高くなったのは固有名詞と名詞両方の質問を強化学習に用いたモデルとなった.

NYT50 における結果を表 3 に示す. 表 3 から, CNN-DM の時と同様に, 強化学習を行ったモデルは baseline と比べ, 正答率, ROUGE 共に向上していることがわかる. また, 同様に固有名詞の質問を強化学習に用いるよりも, 名詞の質問を用いた方が ROUGE が高くなっていることが確認できる. NYT50 データセットでは, 固有名詞と名詞両方の質問を強化学習に用いるよりも,

表 4: 質問数を変化させた場合の結果

平均質問数	Acc_{nnp_nn}	R-1	R-2	R-L
2	30.57	40.55	18.12	37.09
4	30.92	40.68	18.29	37.23
8	30.66	40.68	18.22	37.21
ALL(16.8)	30.95	40.68	18.26	37.22

名詞のみを用いた方が ROUGE が高くなった.

CNN-DM と NYT50 で結果が異なるのは, 原文書に出現する固有名詞の平均数が NYT50 では CNN-DM の約 1.5 倍と多いことから固有名詞を対象とした質問の強化学習が CNN-DM に比べ難しかったため, 名詞の質問のみを使用した方が強化学習がうまくいき, ROUGE が高くなったと推測される.

また CNN-DM データセットにおいて, QA_{nnp} + QA_{nn} の条件で, 質問数を制限した場合の結果を表 4 に示す. 平均質問数は, 学習に使用する一要約に対する質問数の平均を示す. 使用する質問は固有名詞と名詞の質問の中からランダムに選択した. ALL は全ての質問を使用した場合の結果である. 表 4 より, 質問は平均 4 問程度は必要だがそれ以上質問を増やしても ROUGE が向上しない結果となった. これは強化学習に使用する質問が多すぎると, 正答率を向上させるように学習することが難しくなり 平均質問数 8 問や ALL では強化学習がうまくいかなかったと推測される.

5 まとめ

抽象型要約における質問応答を利用した強化学習において, 新たに名詞をマスクした穴埋め問題を質問として用いることを提案した.

まず Scialom ら [3] の結果と同じく, 質問応答におけるスコアが向上するように強化学習を行うことで ROUGE が向上することが確認された. また, CNN-DM と NYT 両方のデータセットにおいて, 従来の固有名詞のみをマスクした穴埋め問題よりも名詞をマスクした穴埋め問題の方が ROUGE の値が高くなることが確認された. 最後に, 強化学習に使用する質問数が増えることでより正確な要約の評価ができ, 強化学習がより効果的になることを期待したが, 最も ROUGE が高くなったのは平均質問数が 4 問の場合で, それ以上は効果が見られなかった.

今後の研究の方針としては, 現時点では人が作成した要約から質問を生成しているが, それを要約対象の原文書から作成し, 教師なしの強化学習を行うことなどが考えられる.

参考文献

[1] C.-Y. Lin, Rouge: A package for auto- matic evaluation of summaries, in: Text summarization branches out: ACL workshop, 2004.

[2] R. Paulus *et al.*, A deep reinforced model for abstractive summarization, in: ICLR, 2018.

[3] T. Scialom *et al.*, Answers Unite! Unsupervised Metrics for Reinforced Summarization Models, in: EMNLP, 2019, pp. 3237–3247.

[4] M. Eyal *et al.*, Question Answering as an Automatic Evaluation Metric for News Article Summarization, in: NAACL, 2019, pp. 3938–3948.

[5] K. M. Hermann *et al.*, Teaching machines to read and comprehend, in: NIPS, Vol. 2015-Janua, 2015, pp. 1693–1701.

[6] A. See *et al.*, Get to the point: Summarization with pointer-generator networks, in: ACL, Vol. 1, 2017, pp. 1073–1083.

[7] S. J. Rennie *et al.*, Self-critical sequence training for image captioning, in: CVPR, Vol. 2017-Janua, 2017, pp. 1179–1195.

[8] Y. Liu, M. Lapata, Text Summarization with Pretrained Encoders, in: EMNLP, 2019, pp. 3721–3731.

[9] G. Durrett *et al.*, Learning-based single-document summarization with compression and anaphoricity constraints, in: ACL, Vol. 4, 2016, pp. 1998–2008.

[10] D. Chen *et al.*, A thorough examination of the CNN/daily mail reading comprehension task, in: ACL, Vol. 4, 2016, pp. 2358–2367.